Cairo University
Faculty Of Engineering
Computer Department

Big Data Report
Crimes in Chicago

Team Members:-

- Ahmed Salah
- Muhammad AbdulKariim
- Maryam Aboulfetouh
- Mai Mahmoud

Table of contents

List Of Figures 3

List Of Tables 3

Project Overview 4

Problem Description 4

Data Set Used 4

Project Modules 5

Analysis and solution of the problem 5

1. Data preprocessing 5

2. Data visualization 5

3. Time Series Analysis 8

3.1 Steps 9

3.1.2 Making the time series stationary 9

3.1.3 Removing Seasonality and Detrending 10

3.1.4 Dividing Data to Training and Testing sets 10

3.1.5 Predicting and Evaluating 10

3.2 Time series analysis of different crime types 12

3.3 Time series analysis of different crime locations 14

4. Geographical distributions of crimes 15

4.1 Steps 15

4.1.1 Extra preprocessing steps 15

4.1.2 Geographical clustering 15

4.1.2.1 Determine optimal number of clusters 15

4.1.2.2 Apply k-means clustering 16

4.1.2.3 Find out locations of clusters centers 17

4.1.3 Heatmap visualization 18

5. Correlations between crime aspects 20

5.1 Steps 20

5.1.1 Extra preprocessing steps 20

5.1.2 Plotting strong correlations 21

Unsuccessful trials 23

Association rules between crime type and variables 23

Future work 24

List Of Figures

- Figure 1: Project Modules.
- Figure 2: Primary Type Frequency.
- Figure 3: District with Crime Types.
- Figure 4: Months with Crime Types.
- Figure 5: Initial Time Series.
- Figure 6: Final Time Series.
- Figure 7: Time Series with First Difference applied.
- Figure 8: Time Series with Predicted Values for the year 2016.
- Figure 9: Time Series with Predicted VS Actual.
- Figure 10: Crammer's V Heatmap.
- Figure 11: Primary type vs Arrests Plot.
- Figure 12: : Primary type vs Domestic Plot.

List Of Tables

- Table 1: Time Series of Different Crime Types.
- Table 2: Time Series of Different Crime Locations.
- Table 3: Number of Clusters.
- Table 4: Scatter Plots.
- Table 5: Cluster Centers.
- Table 6: Crime Heatmaps.

Project Overview

1. Problem Description

Our project aims to answer the following questions to help law enforcement authorities to take necessary safety precautions and better utilize and distribute resources to mitigate crime and protect citizens.

In our project we present the following:

1. Data Visualization relative to crime type
2. Time series forecast of crime rates over the next year (relative to present data).
3. Time series analysis on crimes of different types and in different locations.
4. Geographical visualization of all crimes using clustering; presenting spots with highest crimes rates.
5. Geographical representation of crime types showing increase in rate.
6. Insight onto correlation between different aspects of the crime data.

2. Data Set Used

Crimes in Chicago from 2001 -2017 provided by City of Chicago and the Chicago Police Department

kaggle.com/currie32/crimes-in-chicago

Project Modules

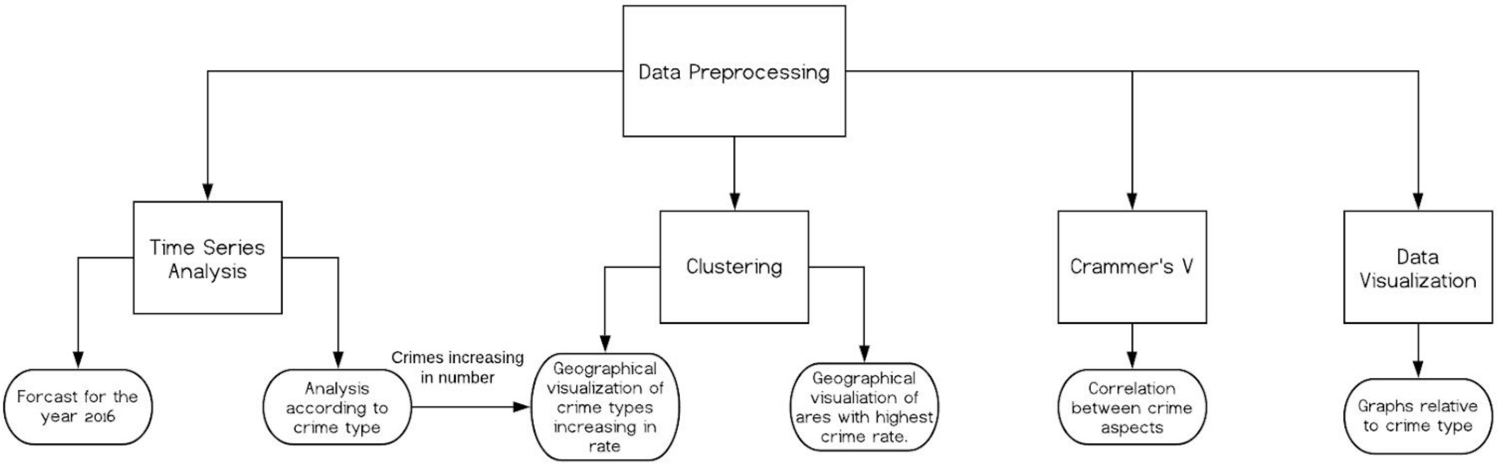

Data Preprocessing
Time Series Analysis
Clustering
Crammer's V
Data Visualization
Forcast for the year 2016
Analysis according to crime type
Crimes increasing in number
Geographical visualization of crime types increasing in rate
Geographical visualiation of ares with highest crime rate.
Correlation between crime aspects
Graphs relative to crime type

Figure 1: Project Modules.

Analysis and solution of the problem

1. Data preprocessing

With a dataset approximately 2 GB in size , some heavy preprocessing was needed to be able to operate on it without R studio crashing.

The following steps were performed before each processing module:

1. Duplicate entries were removed (on batches of data on certain computers).
2. Certain columns were chosen; ones relevant to the process.
3. Null and empty entries were removed.

2. Data visualization

The following figures were extracted to show data composition. Primary types, which state the category of the crime, are of most interest, since, usually law enforcement resources differ with category.

Figure 2 shows the crime frequencies, giving us an idea about crime type distribution in the dataset.

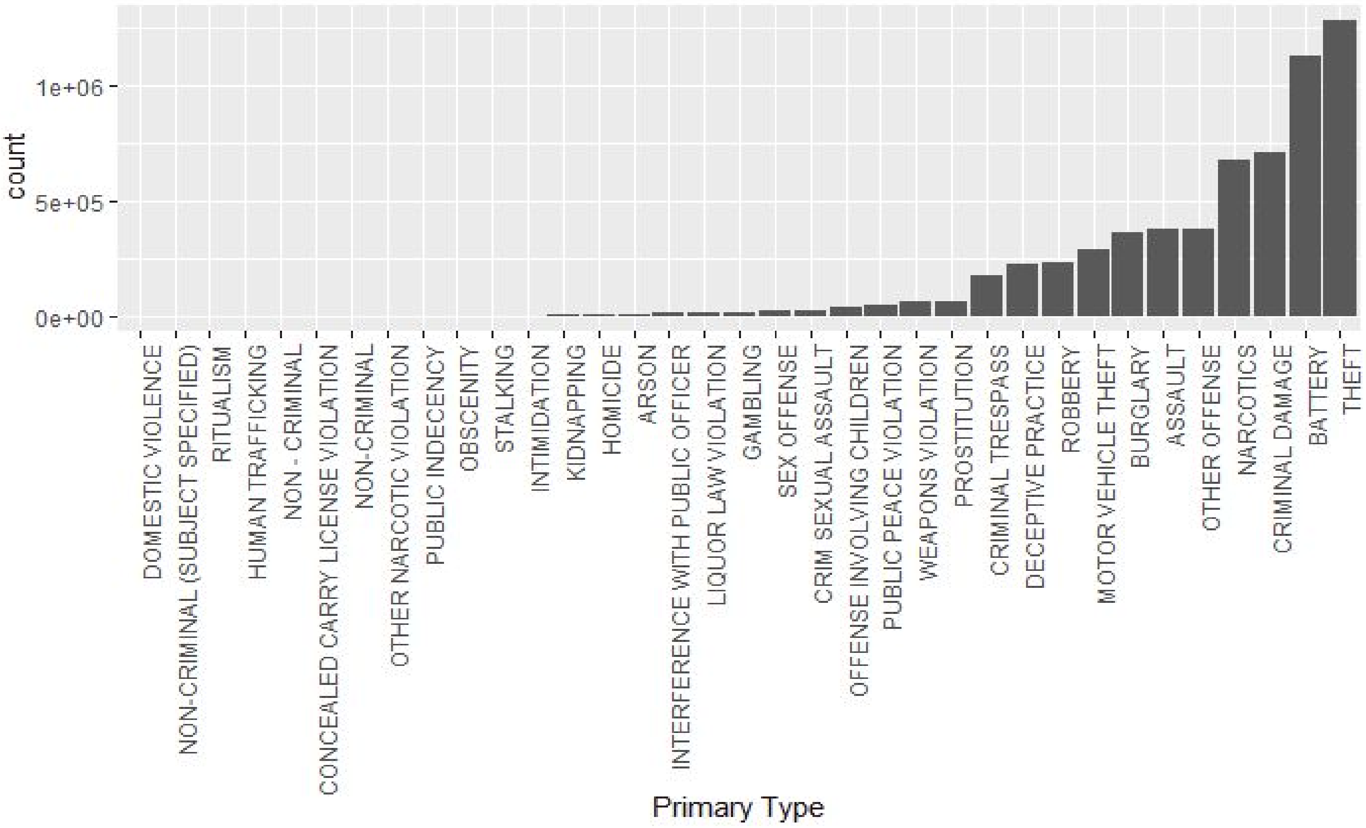


Figure 2: Primary Type Frequency.

Figure 3 shows the crime types in each police district, it is worthy to note that there are districts with very low crime numbers, could be due to an error in recording or just them being new. Also, crimes seem to maintain their distribution across all districts with theft and robbery being most common, so we won't expect a correlation here.

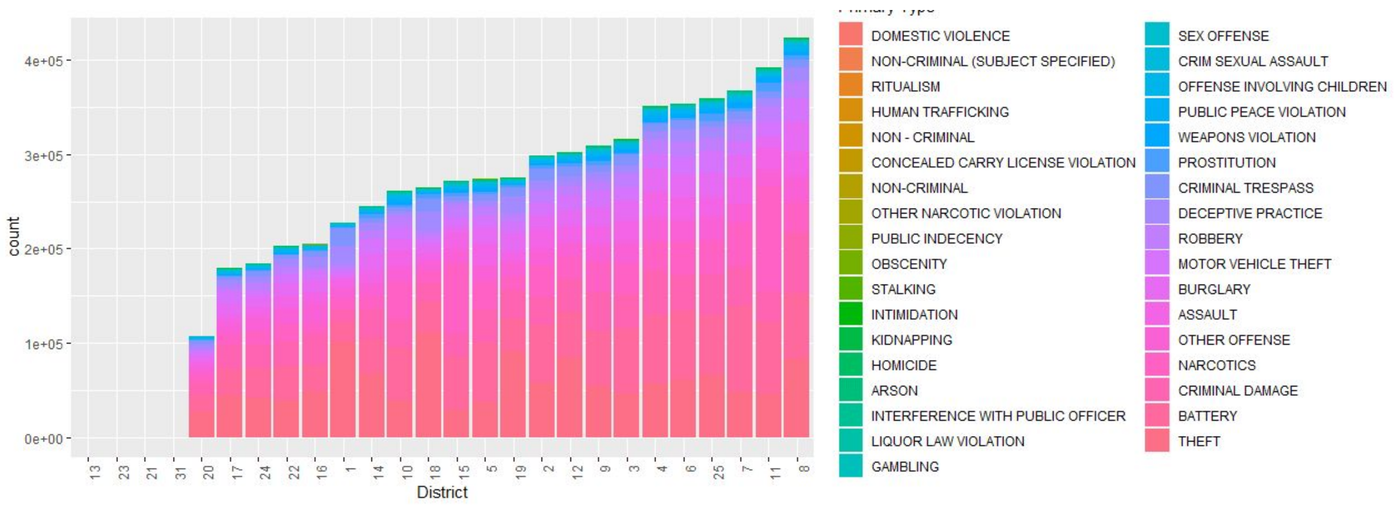


Figure 3: District with Crime Types.

Figure 4 shows monthly variation of the number of occurrences for different types of crimes. **July** has the highest number of crimes.. We can see that Theft, Battery and Criminal damage are the most repeated crimes in July.

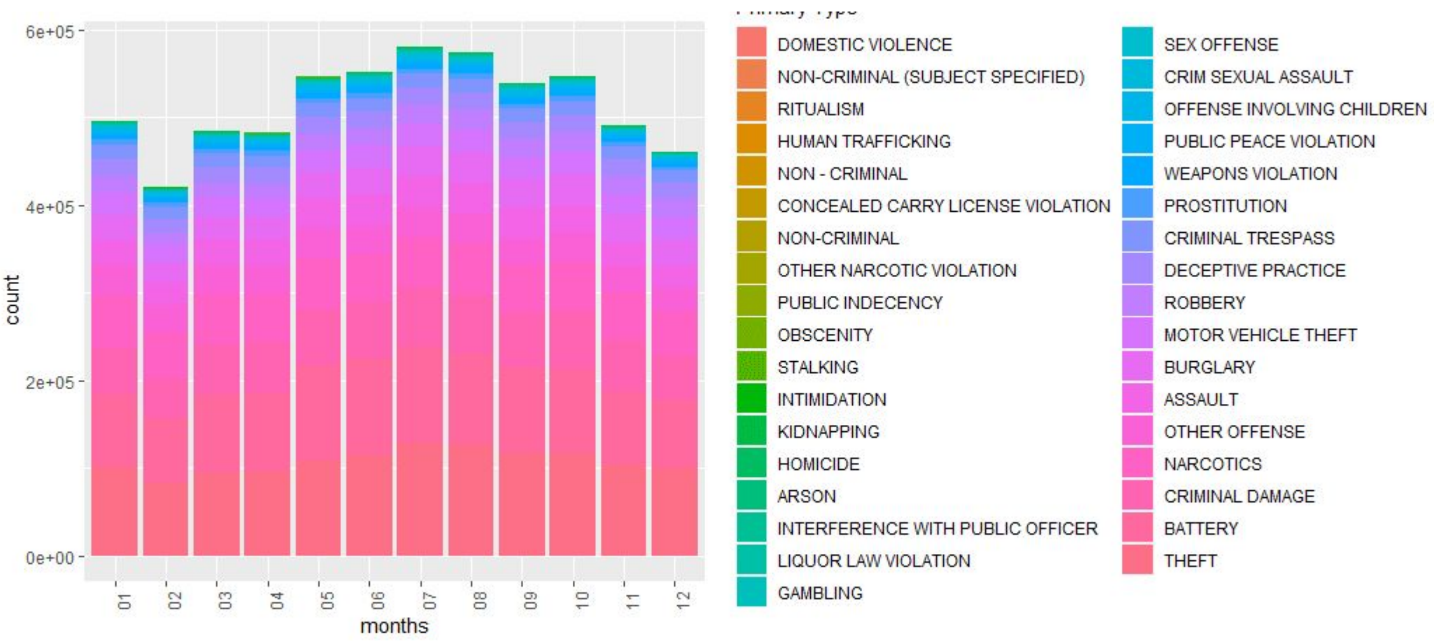


Figure 4: Months with Crime Types.

3. Time Series Analysis

The aim of this process is to produce predictions of the crime rate in Chicago over the next year.
We've plotted the times series of the data dated between 2001 and beginning of 2017.

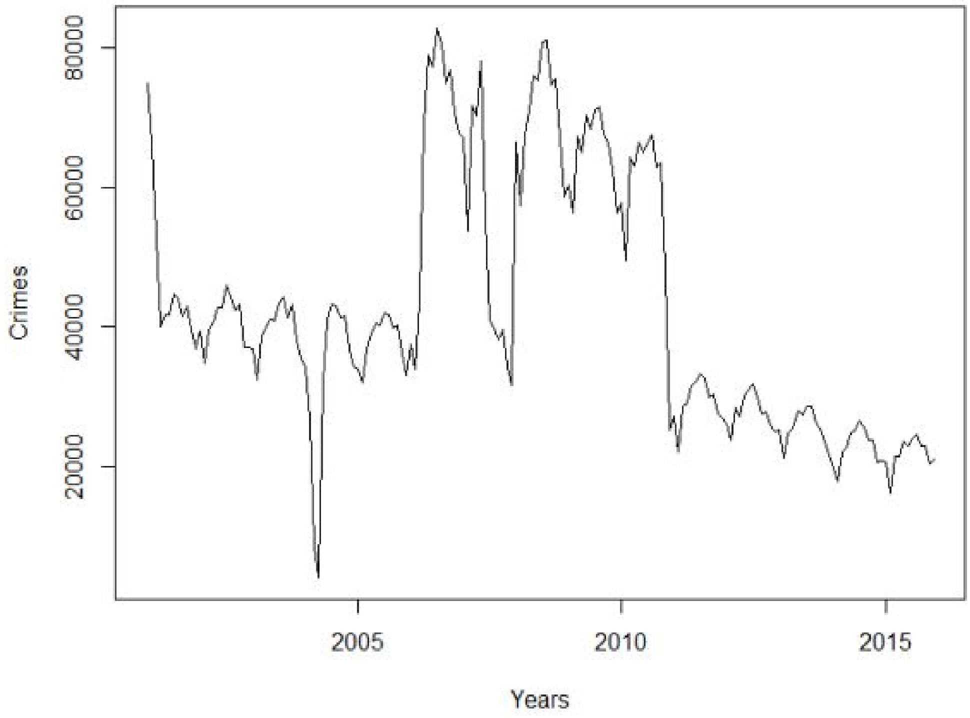


Figure 5: Initial Time Series.

It is evident from *Figure 5* that the data around the year 2004 has an immense deficiency in observations at some point, so we decided to discard it.

It is also apparent from the graph that there is a great increase in observation numbers during the periods between 2006 and 2011, which, upon further inspections turned out to be duplicate records.

Discarding the data prior the year 2005 and removing the duplicate records produced the time series shown in *Figure 6*.

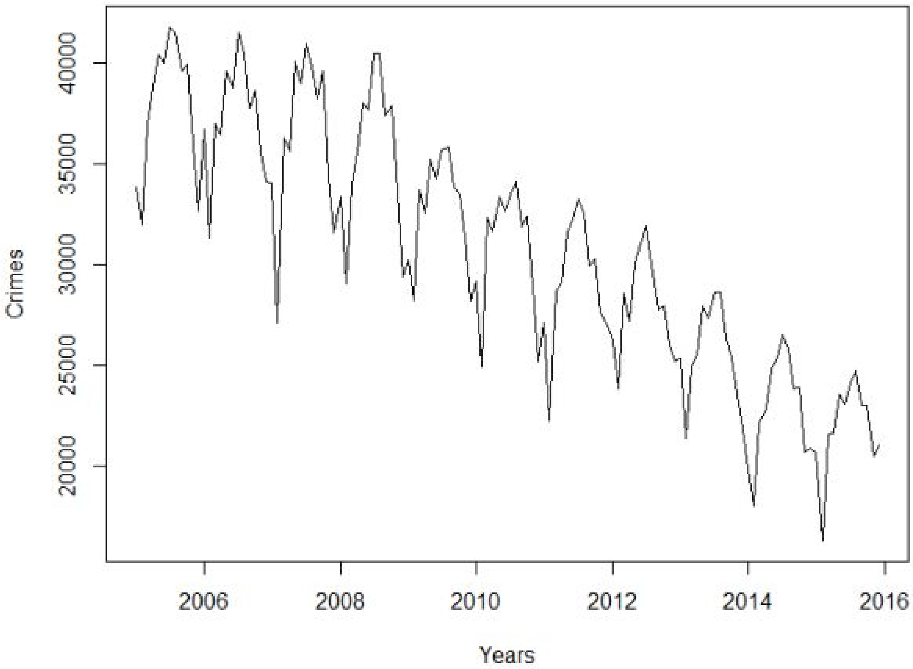


Figure 6: Final Time Series.

3.1 Steps

Moving on to the steps required to produce a Time series:

1. Making it stationary
2. Removing seasonality and detrending

3.1.1 Extra preprocessing steps

Since the Date in the dataset was a string, we needed to format it as “Date” type, so we could plot the Time series.

3.1.2 Making the time series stationary

It is evident from *Figure 6* that there is a “decreasing” linear mean, which could be stabilized using first difference. And stabilize variance using 10log().

Applying the first difference and 10log() produced the following in *Figure 7.*

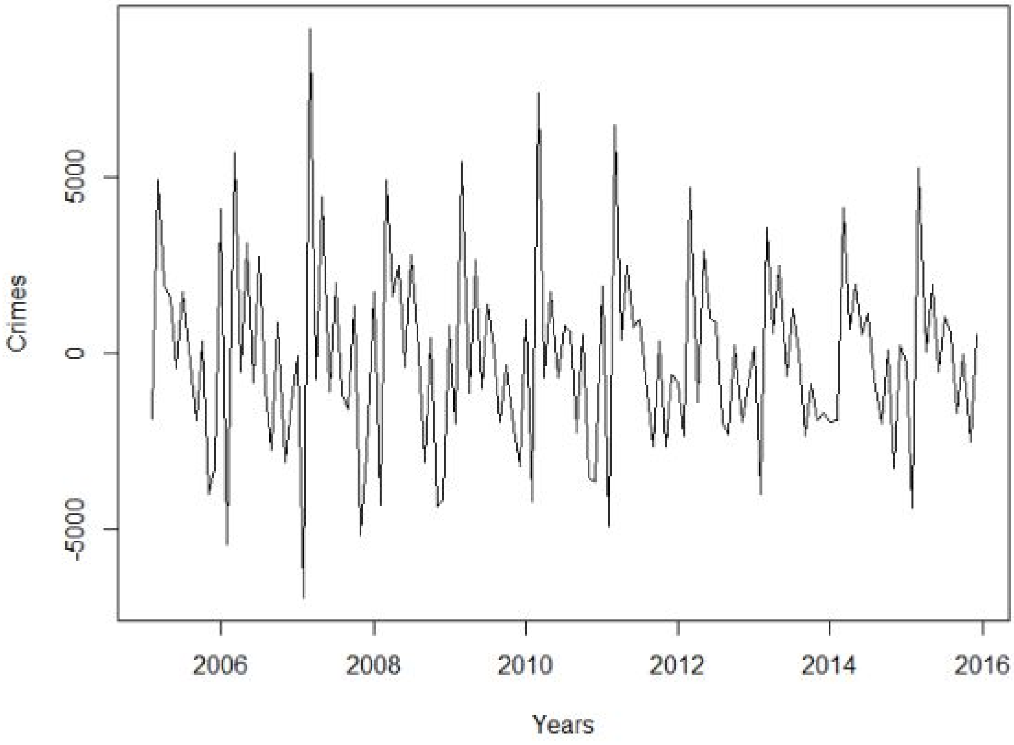


Figure 7: Time Series with First Difference and 10log() applied.

The mean and variance now look constant. Thus, finally stationary.

Note: Actually, the variance seemed constant without the 10log. But performing the 10log produced higher accuracy. See below in 3.1.5.

3.1.3 Removing Seasonality and Detrending

Making the mean constant removed the trend. As for the seasonal adjustment, we'll leave it to ARIMA to take care of. We also decided to not difference the data given to ARIMA as to not have to deal with integrating it before plotting the model predictions.

3.1.4 Dividing Data to Training and Testing sets

It didn't make sense to have a development set because there weren't really any hyper parameters to tune and having a development set would mean we'd have to include it in the model later so we could use the testing set (a time series wouldn't make sense with a missing year in the middle).

We've divided the data into a training set, from 2005 till 2016 and a testing set composing of the year 2016.

3.1.5 Predicting and Evaluating

The predicted values over the year 2016 are shown in *Figure 8*.

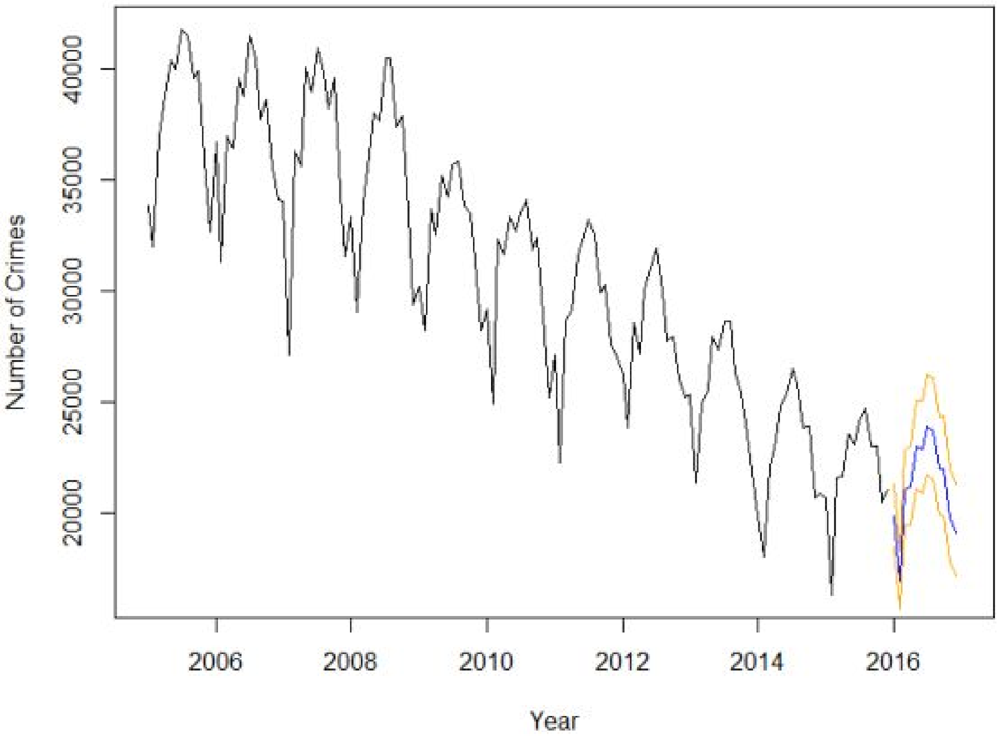


Figure 8: Time Series with Predicted Values for the year 2016.

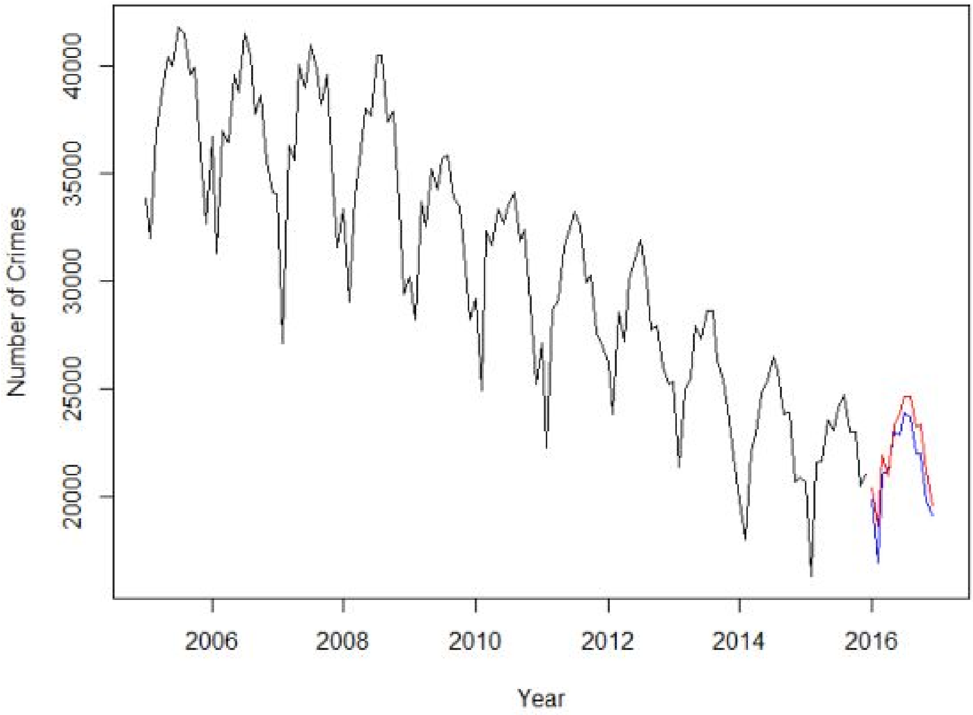


Figure 9: Time Series with Predicted VS Actual.

And the actual vs predicted is shown in *Figure 9*.

Running the accuracy() function in the forecast Library showed the following numbers.

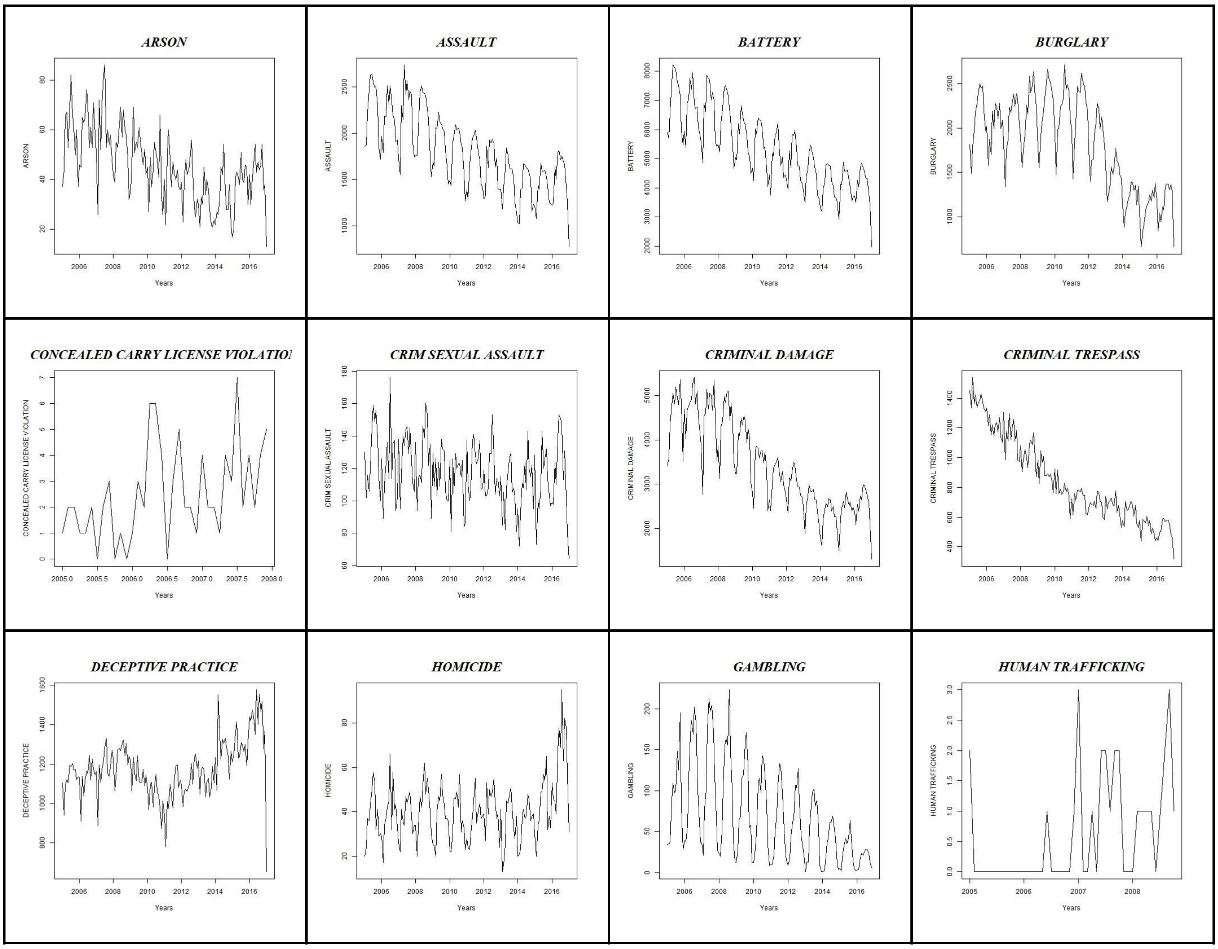
ME	RMSE	MAE	MPE	MAPE	ACF1
-862.385	990.213	888.532	-4.165513	4.289322	0.1680684

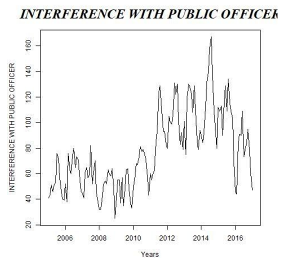
Having a Mean Absolute Percent Error (MAPE) of 4 % is quite decent, given that usually a forecasting approach is deemed successful if MAPE < 10%.

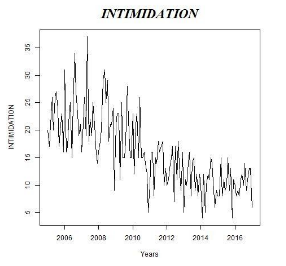
Running accuracy() function on the Time series without 10log, produced a MAPE of 6.89332%, so the difference is quite small.

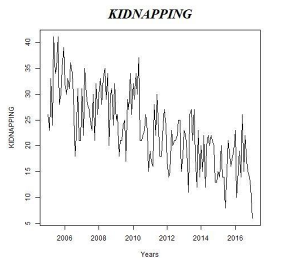
3.2 Time series analysis of different crime types

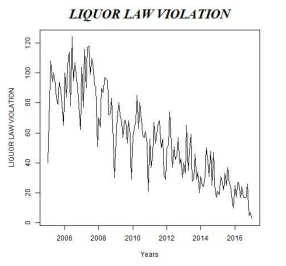
Next we decided to produce a time series for each crime type to help the police department make decisions regarding resource distributions, the following graphs were produced.

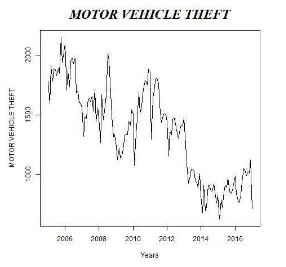


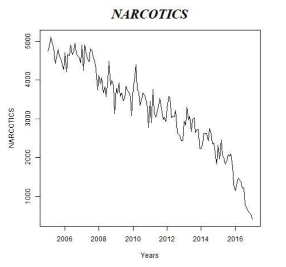

INTERFERENCE WITH PUBLIC OFFICEI
INTERFERENCE WITH PUBLIC OFFICER
20
40
60
80
100
120
140
160
2006
2008
2010
2012
2014
2016
Years

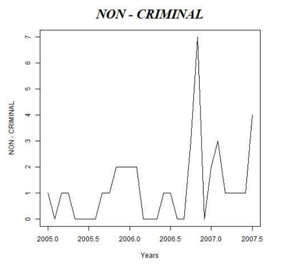

INTIMIDATION
INTIMIDATION
5
10
15
20
25
30
35
2006
2008
2010
2012
2014
2016
Years

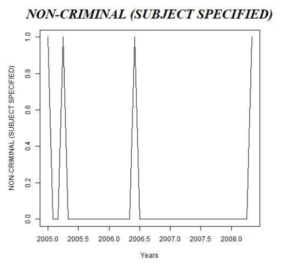

KIDNAPPING
KIDNAPPING
5
10
15
20
25
30
35
40
2006
2008
2010
2012
2014
2016
Years

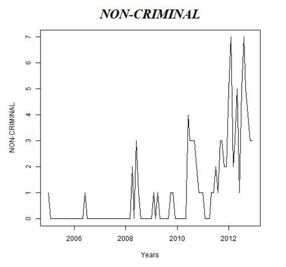

LIQUOR LAW VIOLATION
LIQUOR LAW VIOLATION
0
20
40
60
80
100
120
2006
2008
2010
2012
2014
2016
Years

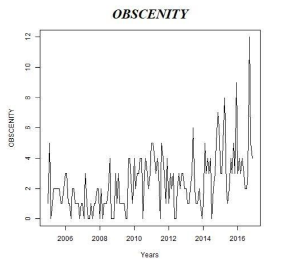

MOTOR VEHICLE THEFT
MOTOR VEHICLE THEFT
1000
1500
2000
2006
2008
2010
2012
2014
2016
Years

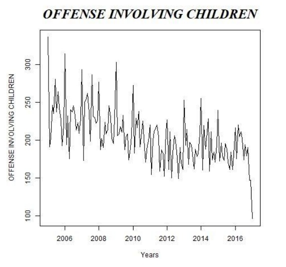

NARCOTICS
NARCOTICS
1000
2000
3000
4000
5000
2006
2008
2010
2012
2014
2016
Years

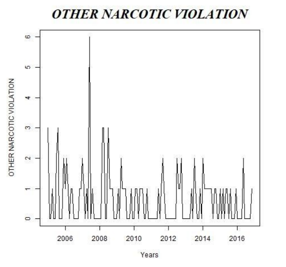

NON - CRIMINAL
NON - CRIMINAL
0
1
2
3
4
5
6
7
2005.0
2005.5
2006.0
2006.5
2007.0
2007.5
Years

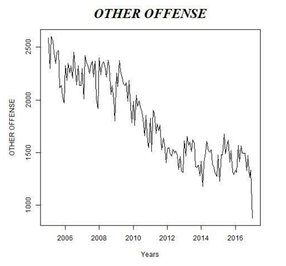

NON-CRIMINAL (SUBJECT SPECIFIED)
NON-CRIMINAL (SUBJECT SPECIFIED)
0.0
0.2
0.4
0.6
0.8
1.0
2005.0
2005.5
2006.0
2006.5
2007.0
2007.5
2008.0
Years

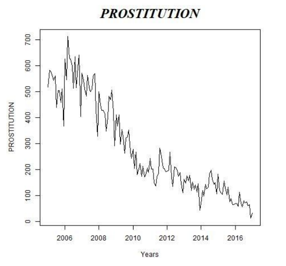

NON-CRIMINAL
NON-CRIMINAL
0
1
2
3
4
5
6
7
2006
2008
2010
2012
Years

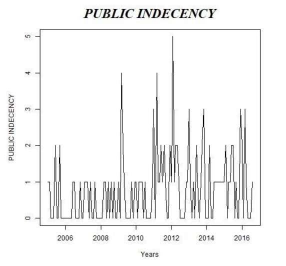

OBSCENITY
OBSCENITY
0
2
4
6
8
10
12
2006
2008
2010
2012
2014
2016
Years

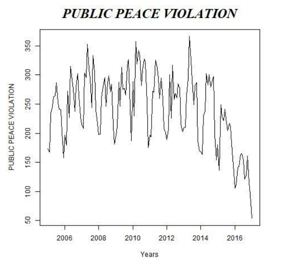

OFFENSE INVOLVING CHILDREN
OFFENSE INVOLVING CHILDREN
100
150
200
250
300
2006
2008
2010
2012
2014
2016
Years

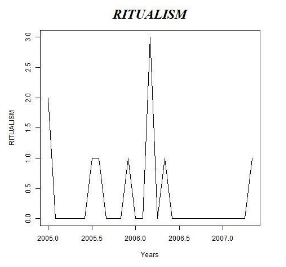

OTHER NARCOTIC VIOLATION
OTHER NARCOTIC VIOLATION
0
1
2
3
4
5
6
2006
2008
2010
2012
2014
2016
Years

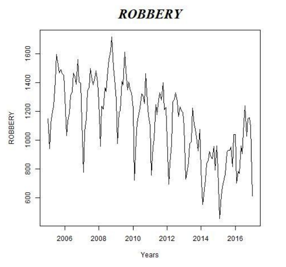

OTHER OFFENSE
OTHER OFFENSE
1000
1500
2000
2500
2006
2008
2010
2012
2014
2016
Years

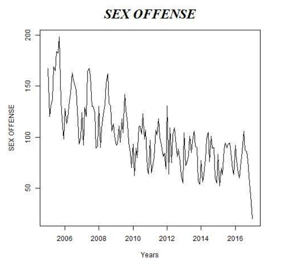

PROSTITUTION
PROSTITUTION
0
100
200
300
400
500
600
700
2006
2008
2010
2012
2014
2016
Years

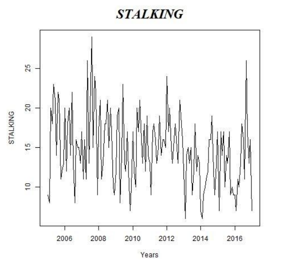

PUBLIC INDECENCY
PUBLIC INDECENCY
0
1
2
3
4
5
2006
2008
2010
2012
2014
2016
Years


PUBLIC PEACE VIOLATION
PUBLIC PEACE VIOLATION
50
100
150
200
250
300
350
2006
2008
2010
2012
2014
2016
Years


RITUALISM
RITUALISM
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2005.0
2005.5
2006.0
2006.5
2007.0
Years


ROBBERY
ROBBERY
600
800
1000
1200
1400
1600
2006
2008
2010
2012
2014
2016
Years


SEX OFFENSE
SEX OFFENSE
50
100
150
200
2006
2008
2010
2012
2014
2016
Years


STALKING
STALKING
10
15
20
25
2006
2008
2010
2012
2014
2016
Years

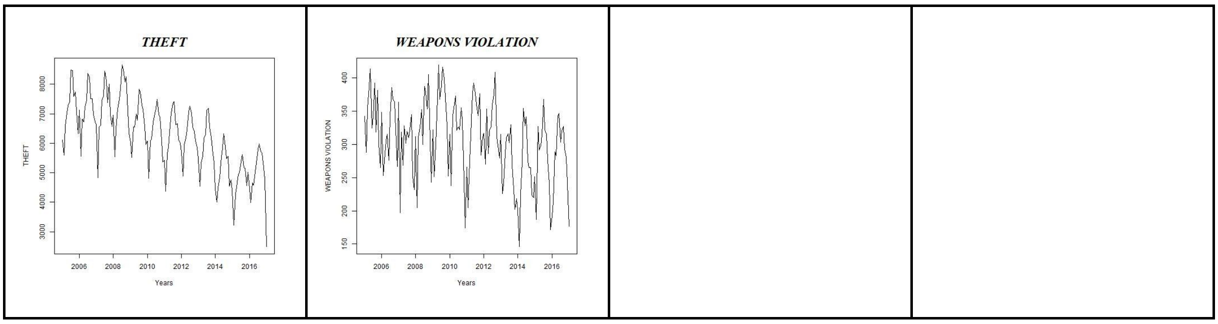


Table 1: Time Series of Different Crime Types.

This draws our attention to the fact that some crime types are actually increasing in the frequency, but didn't show due to their low percentage. This information could be used to help the police to better distribute their resources to fight rising crime numbers.

3.3 Time series analysis of different crime locations

Same logic could be applied to other aspects of crime such as Location (where the crime happened) .
Here are the plots for the 10 most common locations (there were over 100 different locations, so we won't show all plots).

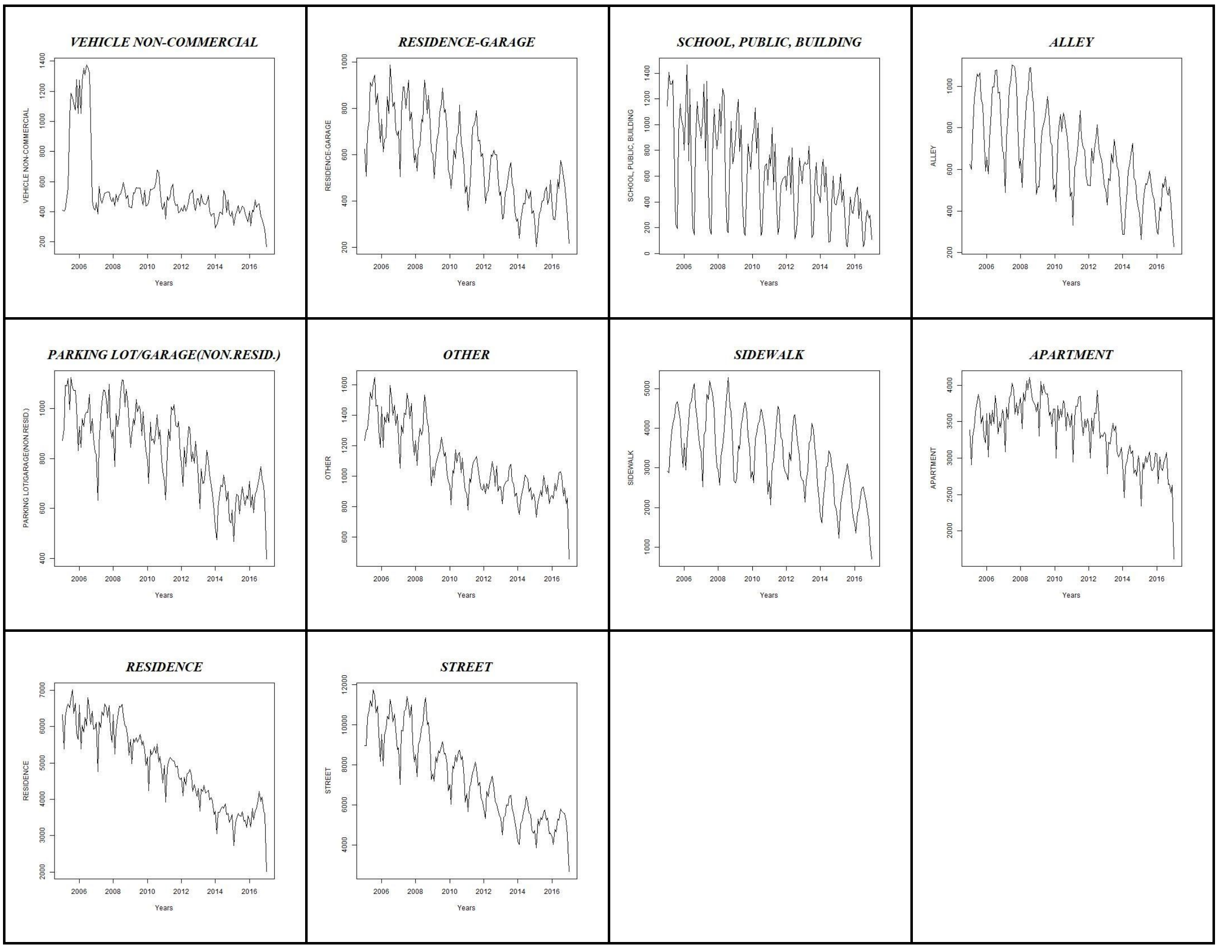


Table 2: Time Series of Different Crime Locations.

4. Geographical distributions of crimes

The aim of this analysis is to show a geographical visualization of the crimes in Chicago and present a useful insight about community areas with the highest crime rates.
Dataset used for this analysis is limited to crimes between 2012-2017. We think these crimes are the most relevant in our geographical analysis.

4.1 Steps

4. Geographical Clustering
    a. Determine optimal number of clusters
    b. Apply k-means clustering
    c. Find out locations of clusters centers
5. Heatmap visualization

4.1.1 Extra preprocessing steps

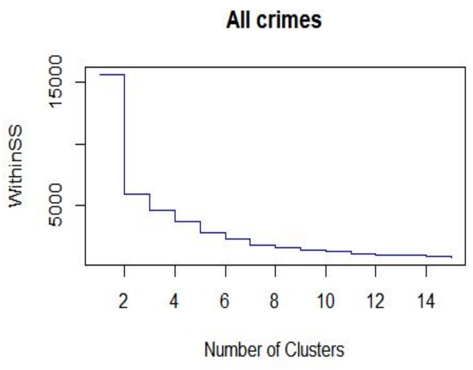
Some crime entries in the dataset were found to be outside Chicago, these entries were removed in the preprocessing phase.

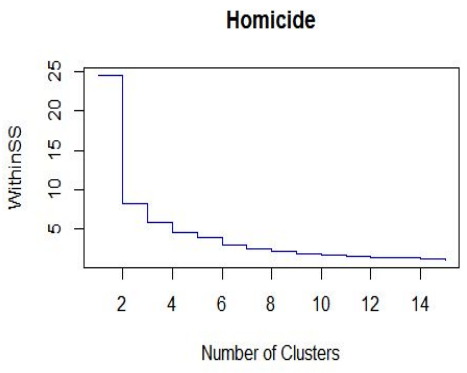
4.1.2 Geographical clustering

To find out locations with the highest crime rates we performed k-means clustering based on the longitude and latitude of crimes.
First, we clustered all the crimes combined then we chose a few increasing crimes based on our Time Series Analysis (Homicide, Robbery and Sexual assault) and clustered them separately.

4.1.2.1 Determine optimal number of clusters

We used the withinSS metric to determine the optimal number of clusters to use in each case. *Table 3* shows the plotting of the number of clusters vs the withinSS.





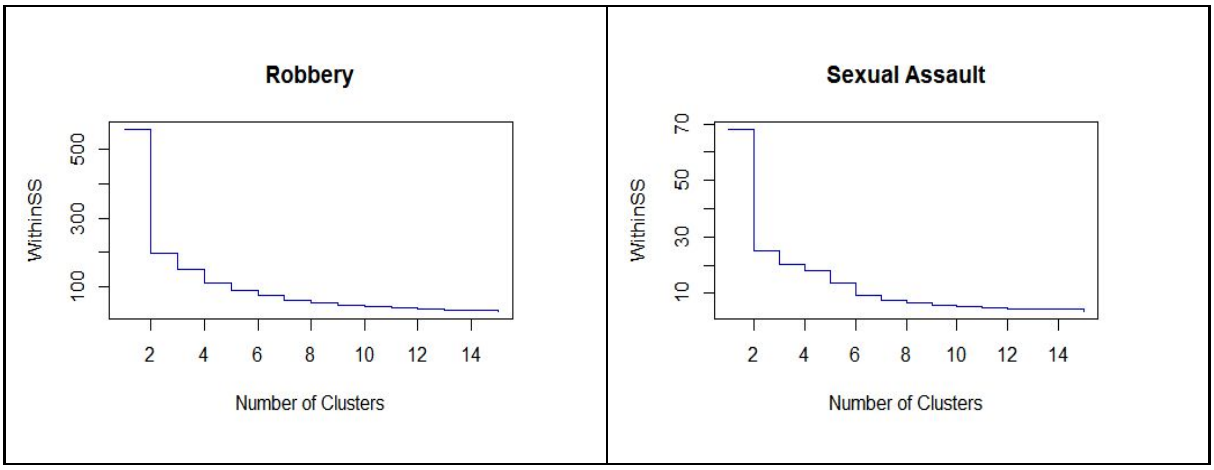
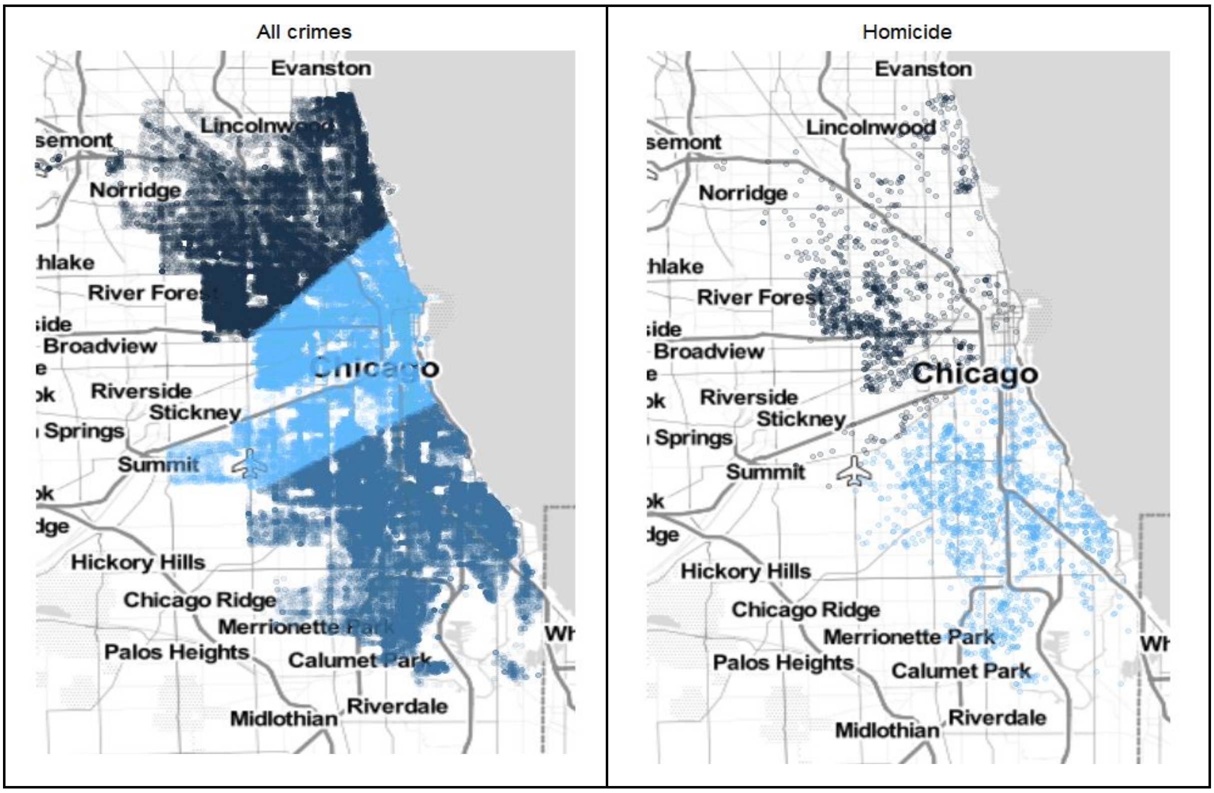


Table 3: Number of Clusters.

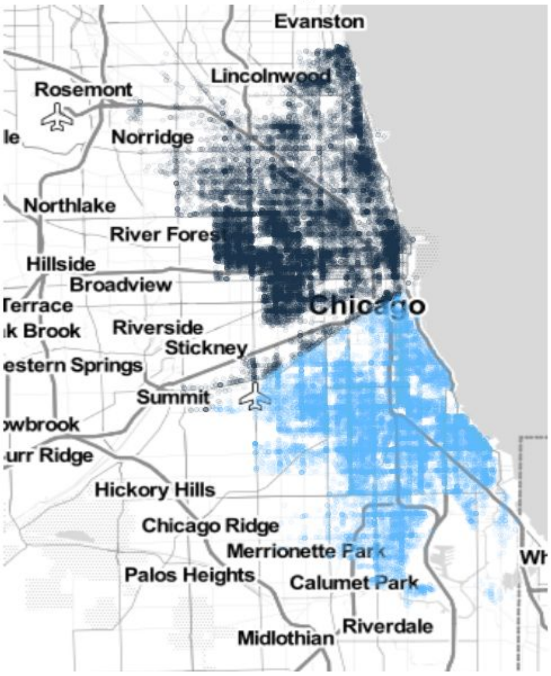
4.1.2.2 Apply k-means clustering

Based on the above figures we concluded that it would be sufficient to use **3** clusters for combined crimes analysis and **2** clusters for individual crimes. *Table 4* shows a scatter plot[1] of crimes colored by cluster and projected on the map of Chicago for each case.



[1] A scatter plot of all crimes was infeasible due to the huge number of points so points were sampled prior to plotting.

Robbery



Sexual assault

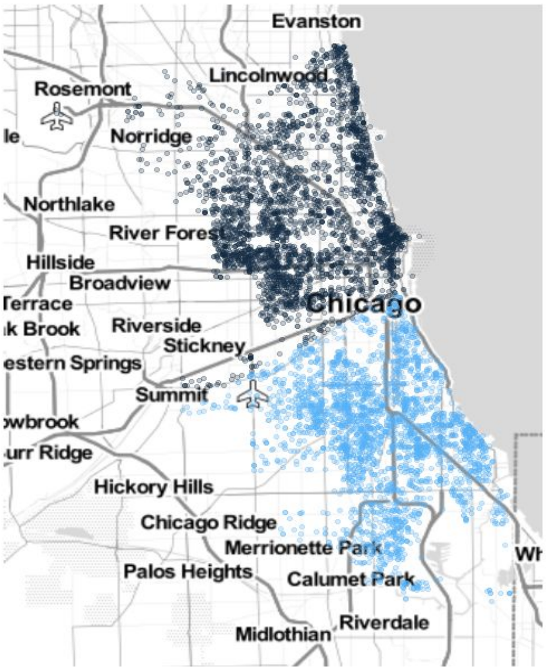


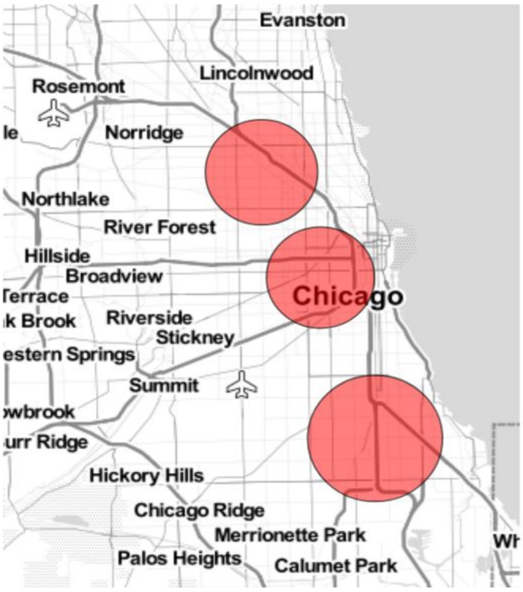
Table 4: Scatter Plots.

4.1.2.3 Find out locations of clusters centers

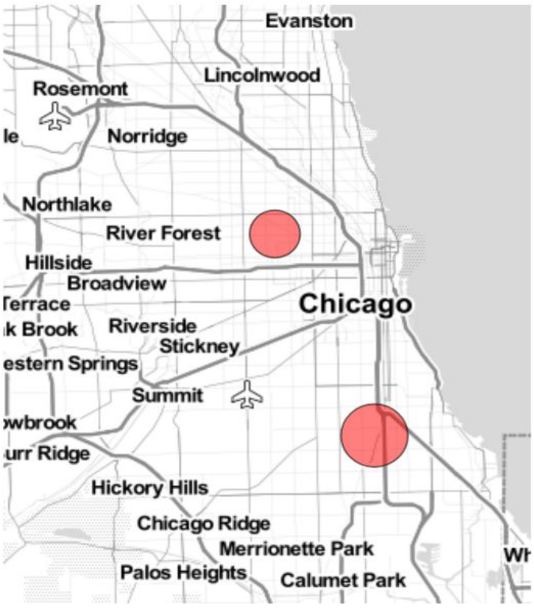
After clustering we projected the cluster centers on the map of Chicago to find out which locations the Chicago police department should look out for the most.

Table 5 shows the cluster centers on the map for each case. Interactive maps showing these location names -which we found out with the help of Google Maps- can be found in the “maps” folder.

All crimes



Homicide



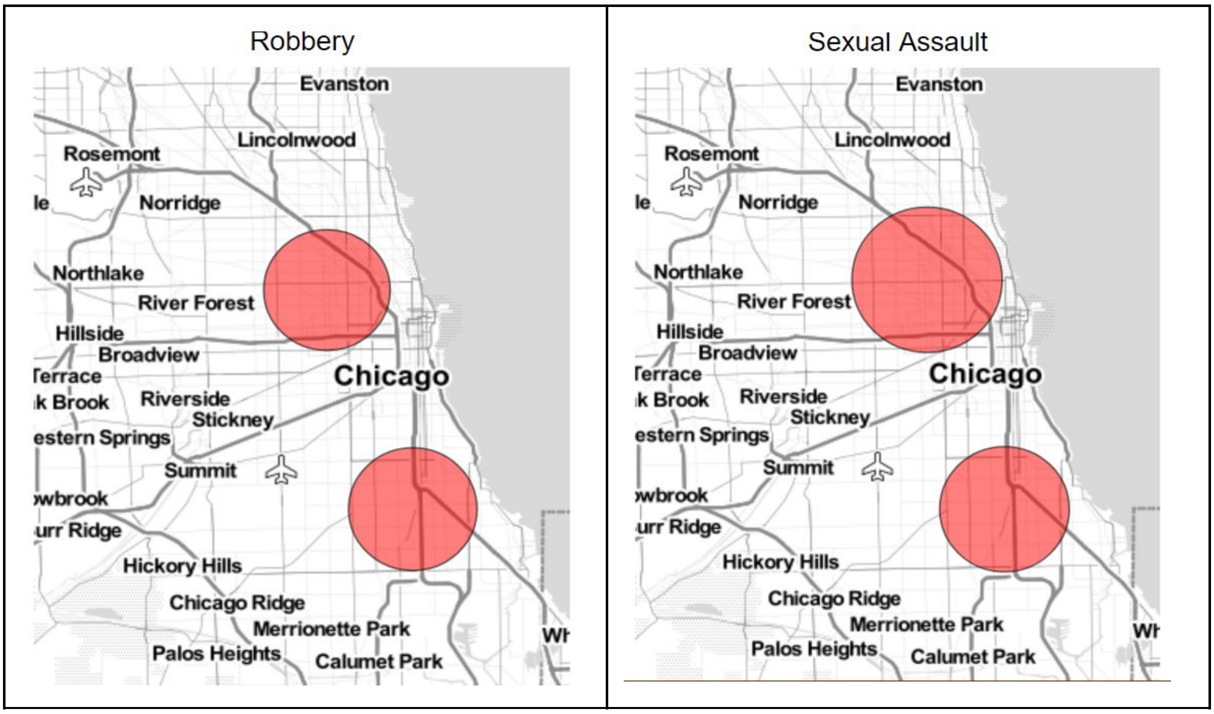
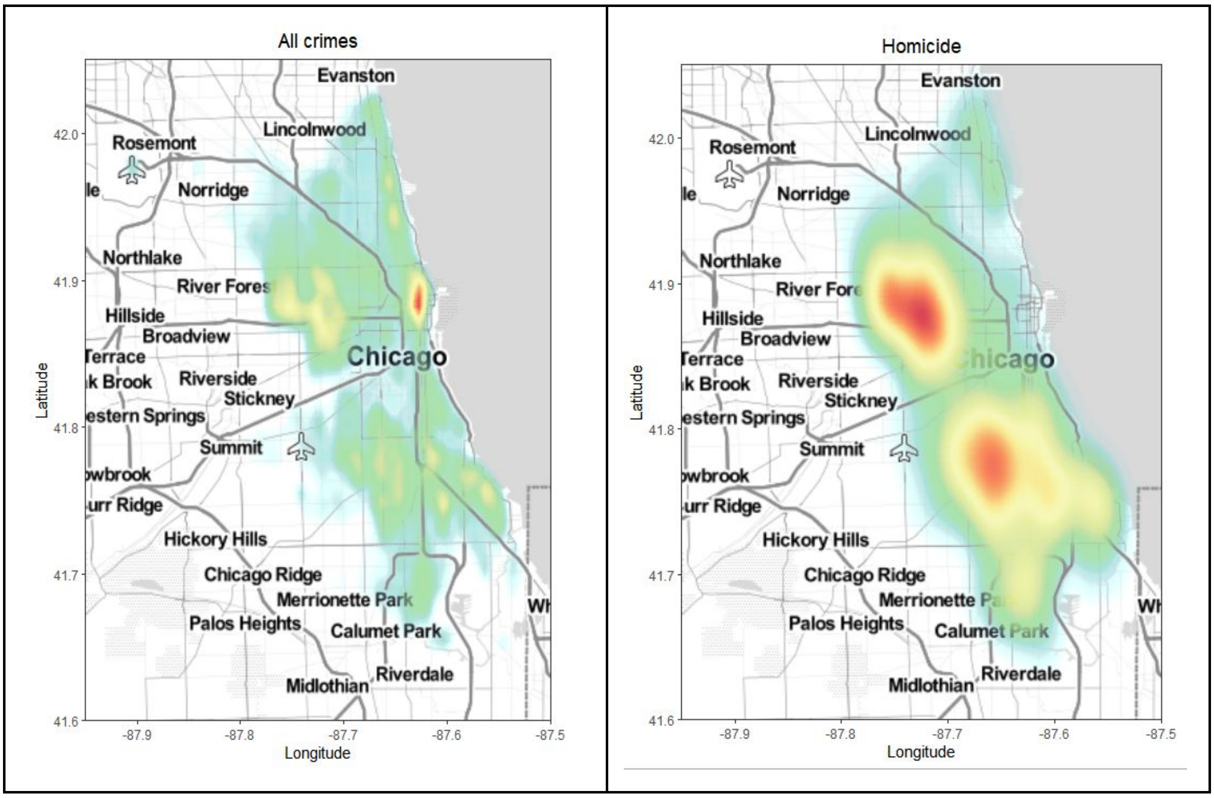


Table 5: Cluster Centers.

4.1.3 Heatmap visualization

For a better visual representation of the crime rate across Chicago we draw a heatmap for all crimes and the three crimes mentioned above. *Table 6* shows the crime heatmaps for each case.



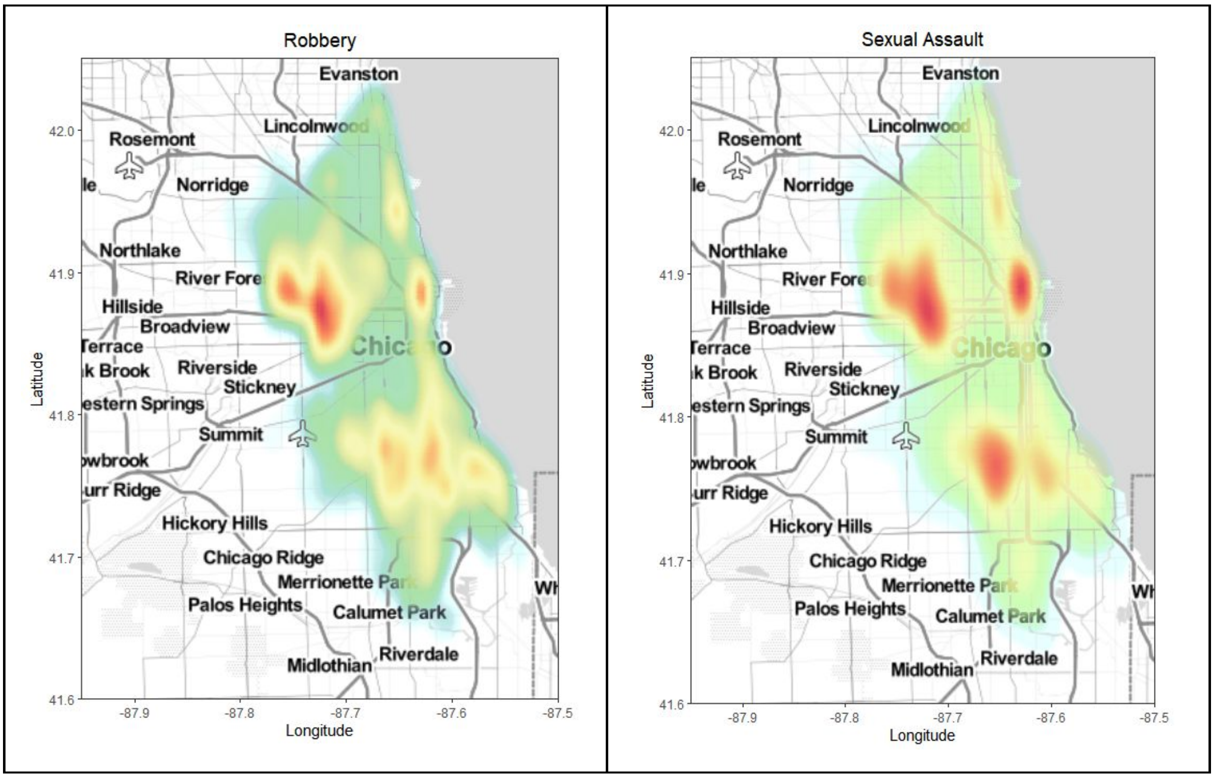


Table 6: Crime Heatmaps.

5. Correlations between crime aspects

The aim of this process is to show any existing correlations between crime aspects of interest. The crime aspects chosen to consider were:

- Date - Date when the incident occurred. This is sometimes a best estimate.
- Primary Type - The primary description of the crime.
- Arrest - Indicates whether an arrest was made.
- Domestic - Indicates whether the incident was domestic-related.
- District - Indicates the police district where the incident occurred.

We've chosen these specifically, because they would produce useful information when it comes to distributing police resources, and focusing on what needs improvement.

5.1 Steps

1. Calculate Crammer's V for all combinations of variable pairs.
2. Plot stacked bar graphs for pairs of high correlation.

5.1.1 Extra preprocessing steps

Since the Date in the dataset was a string, we needed to format it as Date type. Then AM or PM, months, days and seasons were extracted to be used as variables when calculating for correlations.

5.1.2 Calculating Pairwise correlations

Since all the variables chosen are categorical, one way to calculate correlation would have been by using one-hot encoding, and breaking each possible option of each categorical feature to 0-or-1 features. But that would have produced an enormous combination of features.

Instead we decided to use Crammer's V,[2] it is a measure of association between two categorical features. It is based on a nominal variation of Pearson's Chi-Square Test[3]. Links to tutorials on both methods are in the footnote for more detail.

To know whether 2 categorical variables are associated, we first use the chi-square independence test. A p-value produced close to zero means that our variables are very unlikely to be completely unassociated in some dataset.

We then calculate the strength of the association using Crammer's V. The output of Crammer's V is in the range of [0,1], where 0 means no association and 1 is full association.

[2] https://www.spss-tutorials.com/cramers-v-what-and-why/

[3] https://www.spss-tutorials.com/chi-square-independence-test/#assumptions

All variable pairs passed the chi-square independence test, all p-values were much less than 0.05. Which is the common threshold for independence.

Figure 10 shows Crammer's V values between different pairs. The diagonal (correlation 1) was removed.

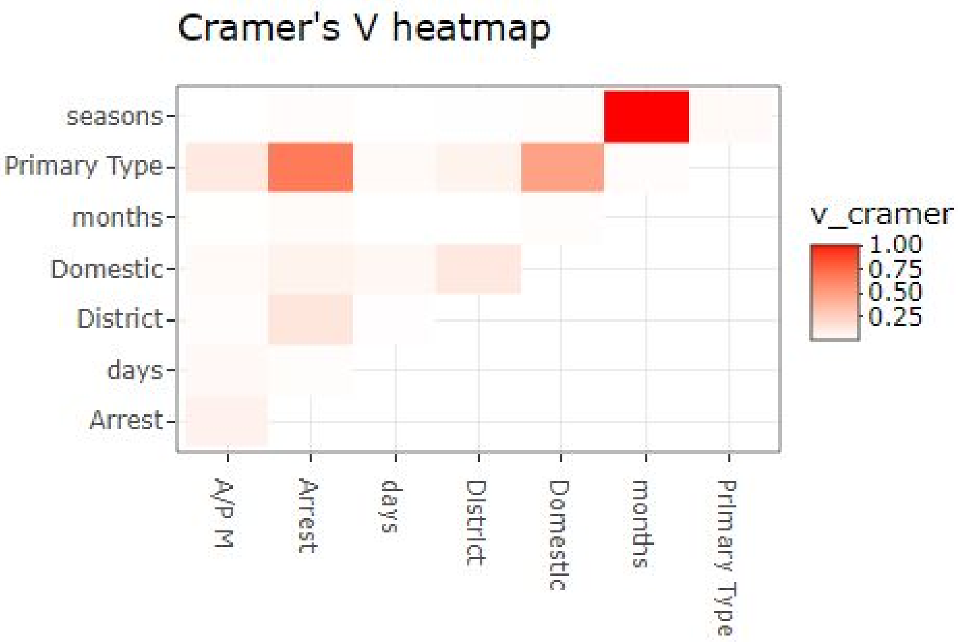


Figure 10: Crammer's V Heatmap.

5.1.2 Plotting strong correlations

As apparent from *Figure 10*, the strongest correlations are between:

- Primary Type and Arrest
- Primary Type and Domestic

Plotting stacked graphs for those relations produced *Figure 11* and *Figure 12*.

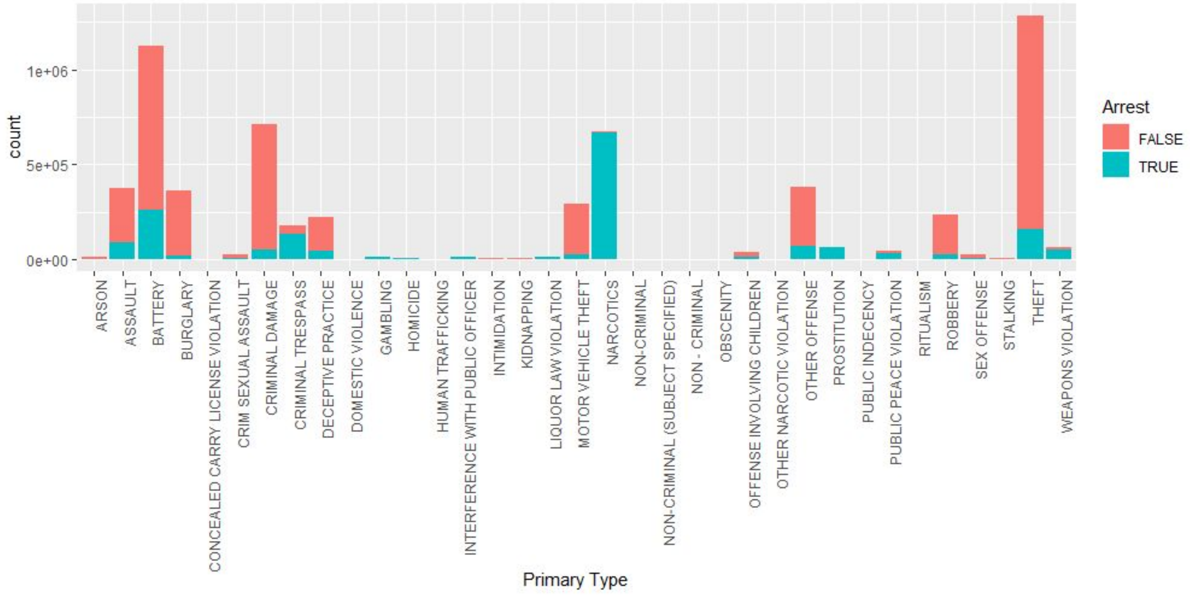


Figure 11: Primary type vs Arrests Plot.

Important information could be drawn from *Figure 11*, for example, numbers of Assault arrests are small relative to failed arrests. Resources need to be redistributed towards crime types with low arrest percentage to reduce the correlation present.

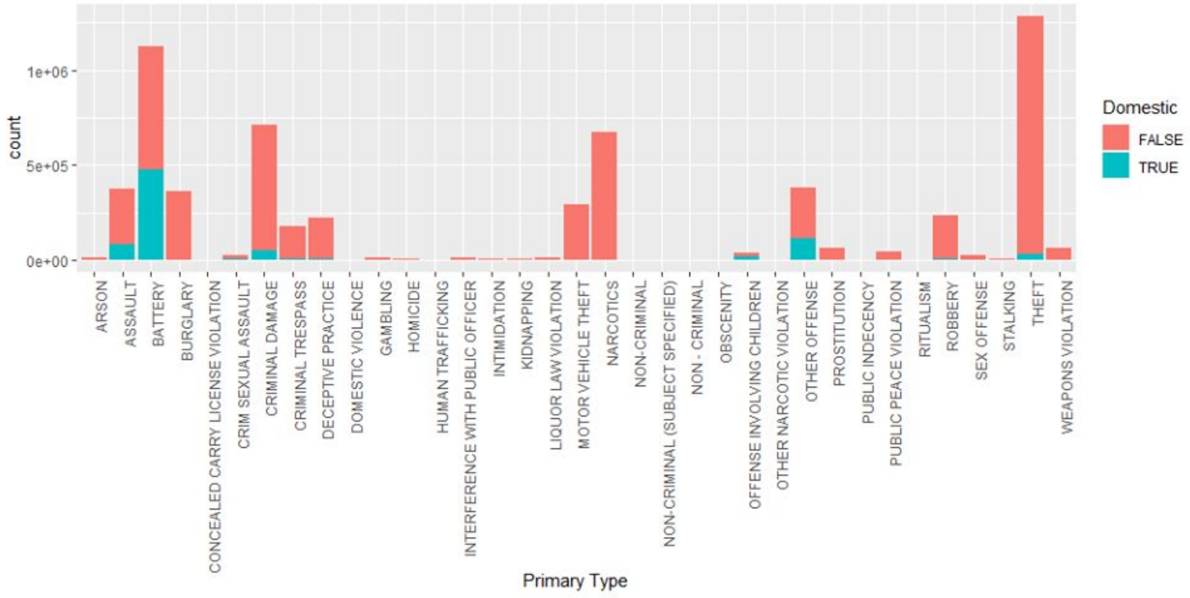


Figure 12: : Primary type vs Domestic Plot.

Figure 12 shows other important insight, such as the fact that Robbery and Theft happen more often outside homes, so security in the streets needs to be strengthened to reduce the correlation.

Unsuccessful trials

Association rules between crime type and variables

Used Apriori algorithm to apply association rule mining techniques to find rules between crime type (on the right hand side of the rule) and these variables on the left hand side of the rule: Weekday - Season - month - district.

Trials with Apriori algorithm:

- Started initially with confident = 0.5 but no rules generated
- Decreased confident value to 0.3, rules generated with bad values of lift (range from 1.1 to 1.3)
- Decreased confident value to 0.01, rules generated with bad values of lift (range from 1.3 to 1.5)

Lift values are a measure of how much the rule is good. Lift value must be high and far from 1 to accept this rule as a meaningful one. The highest rule generated with trails had lift=1.5 which is low. Therefore We concluded that the association rules approach was not suitable to be used in our project.

Future work

Could build a program that would allow the user (in our case, police) to enter a specific crime type and output it’s time series, geographical visualization (of where it is most common) and correlation between its aspects (how different districts dealt with this crime, number of arrests..etc).

In data correlation, we could use Theil’s U, which will provide an asymmetric way to view correlation, it will help us know which variable/crime aspect “drives“ which. Furthermore, relationships between crime rates and day/month/season could be explored (numerical-categorical correlation).